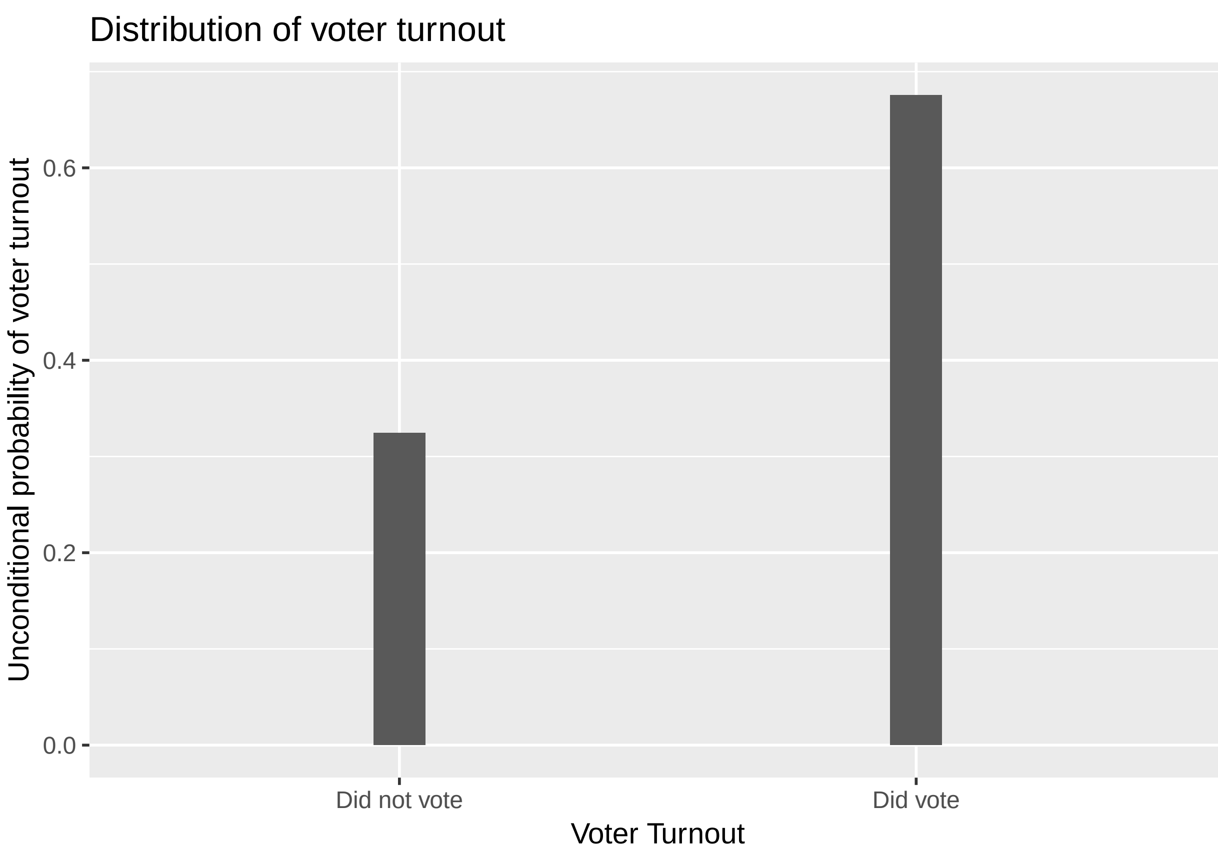
Homework 02: Modeling Voter Turnout

Ellen Hsieh

1.

```r
library(ggplot2)

mh = read.csv('/Users/ellenhsieh/Documents/UChicago/2019 Winter/Modeling/hw02/data/mental_health.csv')

mh = mh %>%
  na.omit #(vote96, mhealth_sum)

mh %>%
  na.omit %>%
    mutate(vote96 = factor(vote96, levels = 0:1, labels = c("Did not vote", "Did vote"))) %>%
    ggplot(aes(vote96)) +
    geom_bar(aes(y = (..count..)/sum(..count..)), width = 0.1) +
    labs(title = "Distribution of voter turnout",
         x = "Voter Turnout",
         y = "Unconditional probability of voter turnout")
```



2.

a.

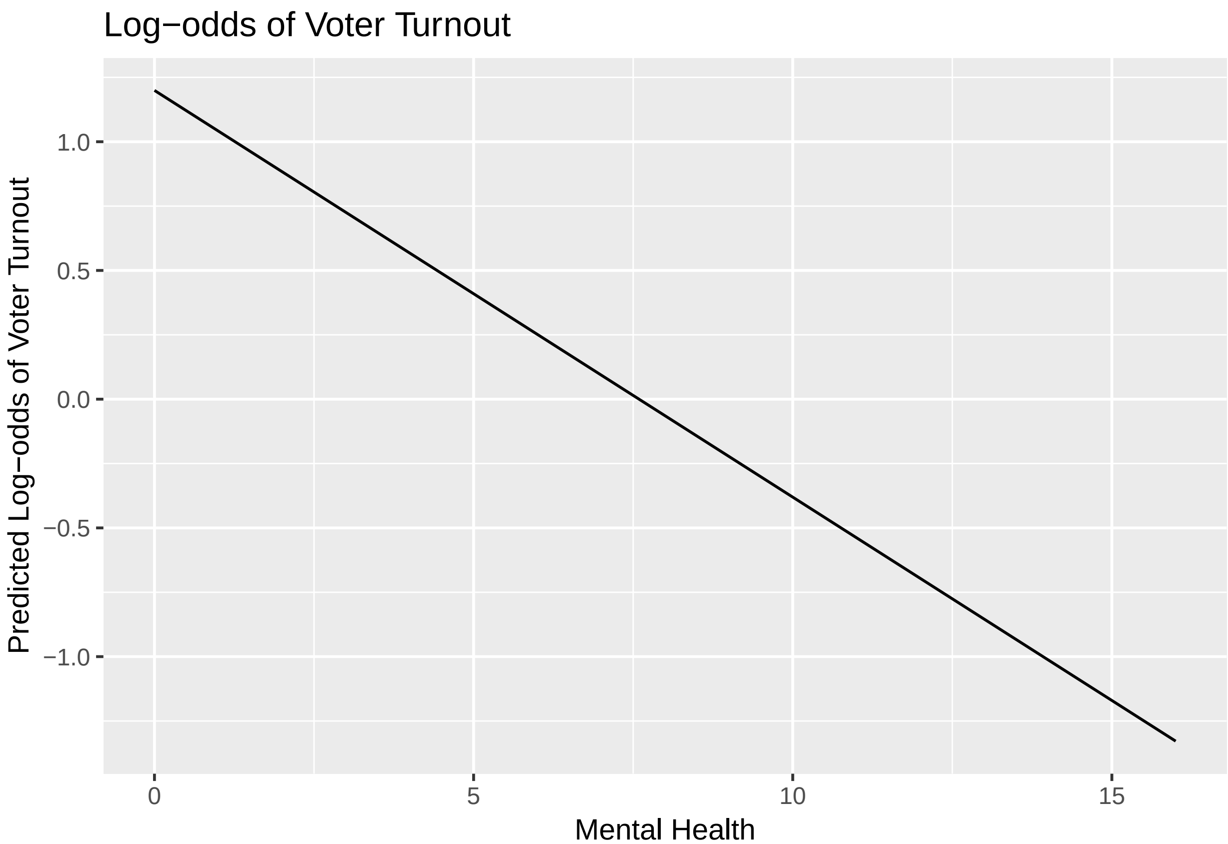
```
vote_mh = glm(vote96 ~ mhealth_sum, data = mh, family = binomial)
summary(vote_mh)
```

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = binomial, data = mh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7105  -1.2859   0.7258   0.8318   1.7682
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.19943    0.09191  13.050  < 2e-16 ***
## mhealth_sum -0.15798    0.02157  -7.324  2.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1468.3  on 1164  degrees of freedom
## Residual deviance: 1411.8  on 1163  degrees of freedom
## AIC: 1415.8
##
## Number of Fisher Scoring iterations: 4
```

Yes, the relationsihp between voter turnout and mental health is statistically significant and negative.

b.

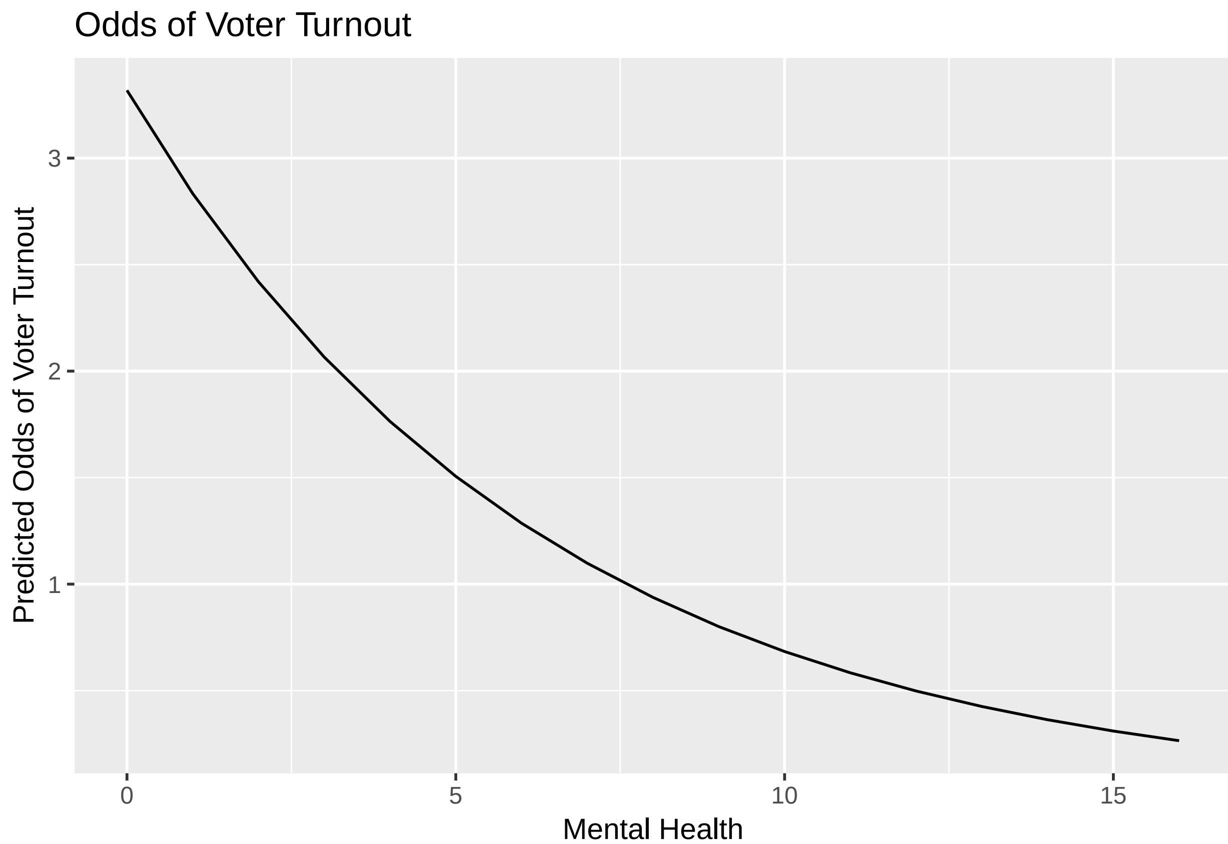
```
new = data.frame(mhealth_sum = 0:16) %>%
  mutate(pred = predict(vote_mh, data.frame(mhealth_sum = 0:16), interval = "prediction"))

ggplot(mapping = aes(new$mhealth_sum, new$pred)) + geom_line() +
labs(title = "Log-odds of Voter Turnout ",
        x = "Mental Health",
        y = "Predicted Log-odds of Voter Turnout")
```



When the point increases by one, the log-odds of voter turnout decrease by 0.158.

c.

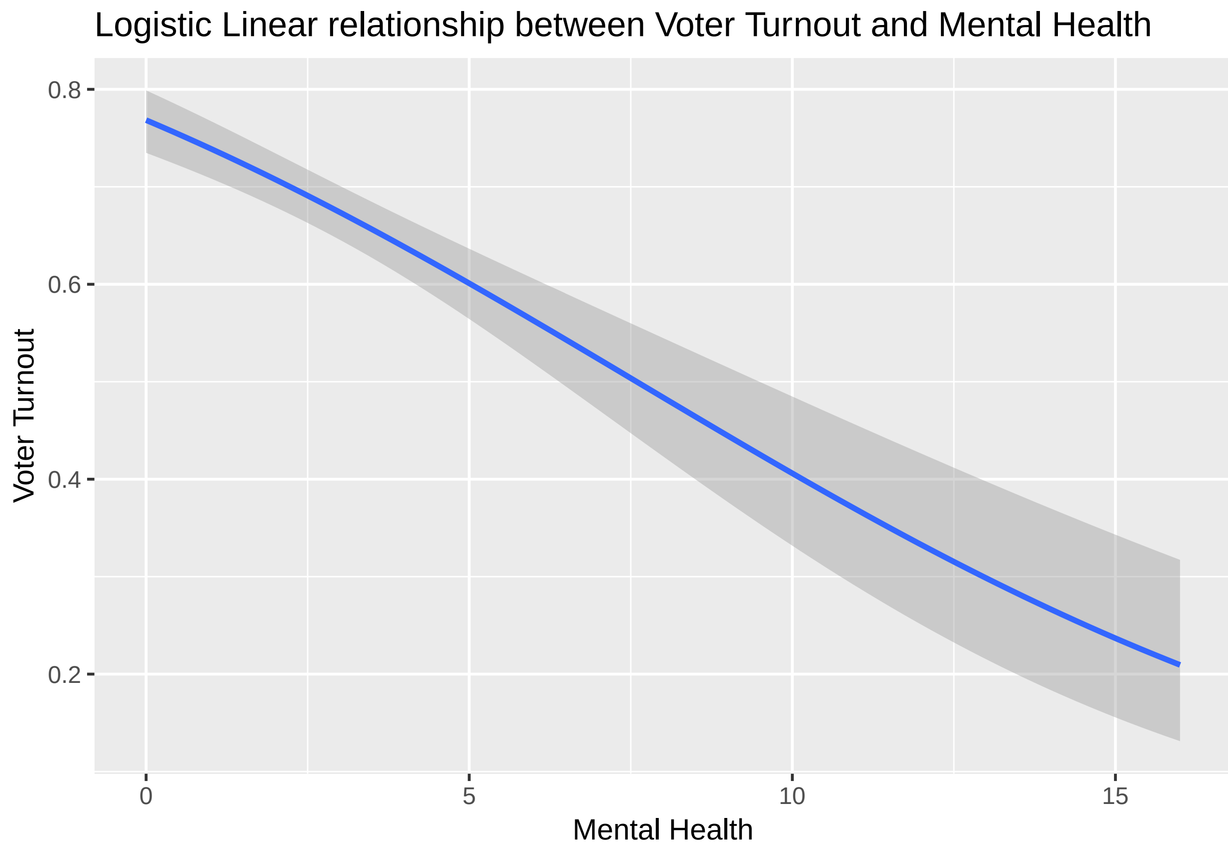
```
ggplot(mapping = aes(new$mhealth_sum, exp(new$pred))) +
  geom_line() +
  labs(title = "Odds of Voter Turnout",
        x = "Mental Health",
        y = "Predicted Odds of Voter Turnout")
```



When the point increases, the odds of voter turnout decreases by $exp(0.158 * beta_1)$..

d.

```
ggplot(mh, aes(x = mhealth_sum, y = vote96)) +
  stat_smooth(method = "glm", method.args = list(family = "binomial")) + labs(title = "Logistic Linear 
        x = "Mental Health",
        y = "Voter Turnout")
```



When the point increases in depression, the probablity of the voter turnout decreases.

e.

```
vote_mh_fit = mh %>%
  add_predictions(vote_mh) %>%
  mutate(pred = logit2prob(pred),
         pred = as.numeric(pred > .5))

mean(vote_mh_fit$vote96 == vote_mh_fit$pred, na.rm = TRUE)
```

```
## [1] 0.6824034
```

68.2% of the predictions based on mental health are correct.Therfore, this model is quite good to predict the voter turnout result.

3

```
vote_model = glm(vote96 ~ mhealth_sum + age + educ, data = mh, family = binomial)
summary(vote_model)
```

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum + age + educ, family = binomial,
##     data = mh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.5592  -1.0412   0.5495   0.8437   2.1330
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.11425    0.48855  -8.421  < 2e-16 ***
## mhealth_sum -0.10719    0.02307  -4.647 3.37e-06 ***
## age          0.04383    0.00484   9.056  < 2e-16 ***
## educ         0.24848    0.02782   8.930  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1468.3  on 1164  degrees of freedom
## Residual deviance: 1260.7  on 1161  degrees of freedom
## AIC: 1268.7
##
## Number of Fisher Scoring iterations: 4
```

This new model contains the parameters of mental health, age and education of the voters. All the new predictors also appears strong relationship with voter turnout. Only the meatal health has negative relationship. Adding more predictors might increase the accuracy of the model if those predictors are significantly related to the response.